THE STATE OF THE FIELD

Qualitative Analyses of Text Complexity

ABSTRACT

The purpose of this article is to understand the function, logic, and impact of qualitative systems for analyzing text complexity, focusing on their benefits and imperfections. We identified two primary functions for their use: (*a*) to match texts to reader ability so that readers read books that are within their grasp, and (*b*) to unearth, and then scaffold, those features of specific texts that are likely to present challenges for readers of differing abilities. We examine three approaches to qualitative text analysis (text-leveling systems, rubric and exemplar systems, and text-mapping systems) relative to these functions. We conclude by strongly advocating the use of qualitative systems, if only to prevent the unchecked use of quantitative approaches from promoting invalid applications of text complexity. In the same breath, we raise a set of vexing issues that the field must address if these approaches are to be used with confidence.

P. David Pearson
UNIVERSITY OF CALIFORNIA, BERKELEY

Elfrieda H. Hiebert
TEXTPROJECT AND UNIVERSITY OF CALIFORNIA, SANTA CRUZ

If the essence of a qualitative system is the use of human judgment, then qualitative systems are not new as tools to help educators determine appropriate texts for use in instruction; they have been in use, in one form or another, for at least 130 years (e.g., Sherman, 1893) as a way of gauging our assessment of the difficulty students are likely to face when reading particular texts. What is new in the current renaissance of qualitative analysis is the deliberate use of qualitative systems as a policy tool, alongside quantitative systems, to shape the reading diet of America's students (National Governors Association [NGA] Center for Best Practices & Coun-

AQ: 1

 0013-5984/2014/11502-0002 $10.00

cil of Chief State School Officers [CCSSO], Appendix A, 2010). The use of tools to control our subjectivity in making judgments (templates, rubrics, and prototypes/exemplars) is a fundamental tool of decision making in many human endeavors, including performance evaluations in the workplace, Olympic sports, and the quality of writing completed by students. It is no different when it comes to matching books to readers. Clearly teachers, librarians, parents, teacher educators—as well as children (recall the five-finger rule for judging difficulty)—have used templates or prototypes for choosing "just right" books for generations. Writers such as Trelease (2006) identify recommendations for particular grade levels, as does Hirsch (e.g., Hirsch, 2005). Lists of recommendations for students reading at different grade levels can be found at a variety of websites.[1]

Whether texts designated by these means provide too much challenge or not enough for particular groups of students is uncertain. To our knowledge, no one has conducted a direct validation of any of these leveling systems to determine whether the texts assigned to a level provide just the right challenge for students judged (or more likely assumed) to be reading at that level. In large measure, those who create and implement these systems are more likely to use anecdotal classroom reports of their success in matching students to books than any sort of careful analysis of student reading performance.

To operationalize any system of human judgment that aspires to match books to students, two estimates are needed: (*a*) an estimate of the level (often operationalized as a grade level) at which a real or prototypic student can read, and (*b*) an estimate, hopefully on the same scale as the student score, of the level of difficulty of a large number of books. Find the level of the readers and let them select from books judged to be at their reading levels—that's the logic.

This article is about systems for scaling books, so we will not dwell on systems for determining the level at which students can read, either on their own or with teacher and peer support, except to say that there is a long and complicated literature on the topic, mostly conducted in the spirit of validating and rationalizing informal reading inventories (Betts, 1946; Pikulski & Shanahan, 1982) and their commercial counterparts (e.g., Beaver, 2003; Leslie & Caldwell, 2010). Suffice it to say that identifying an individual student's true reading level is much more complicated than can be inferred from an informal assessment (see Valencia, Wixson, & Pearson, 2014, in this issue, for an account of why any individual's reading level may not be as stable an estimate as is assumed by informal assessments).

The business of matching texts to readers is not the exclusive purview of qualitative approaches to measuring text difficulty. The quantitative scaling of books and the matching of books to readers are addressed in two articles in this special issue (Cunningham & Mesmer, 2014, and Williamson, Fitzgerald, & Stenner, 2014). The focus of this article is on qualitative approaches for scaling texts, most often to allow teachers to match them to their students' current reading capacities, but also to provide teachers with insights that might help them in teaching particular texts.

The three qualitative systems reviewed in this section distinguish themselves from the informal approaches to text leveling that have emerged from grass roots efforts (e.g., Rog & Burton, 2001) in that they are more systematic in describing—and/or analyzing, and/or validating—their criteria and procedures. All three approaches have been described in published documents, although, as we indicated, precious little is known about the validity of the text assignments in relation to measures of

student reading (e.g., accuracy, fluency, or comprehension) or to teachers' efficacy in providing appropriate instruction. The first approach—text leveling (TL)—is used extensively in school contexts and is described in the pedagogical literature (e.g., Fountas & Pinnell, 1996, 2009; Peterson, 1991). The second approach—rubrics plus exemplars (R+E)—is the one promoted within the CCSS and used in several prior efforts (e.g., ACT, 2006). A third approach—text maps (TM)—is used by the National Assessment of Educational Progress (NAEP; American Institutes for Research [AIR], 2008) and was in use in several state assessments in the 1980s and 1990s to determine the critical content of a text (Valencia, Pearson, Peters, & Wixson, 1989; Wixson, Peters, Weber, & Roeber, 1987).

AQ: 2

The TL and R+E systems are similar in that both rely on two elements: (*a*) the use of criteria for describing and rating text complexity and (*b*) the use of exemplar texts to "anchor" what is expected of readers at different levels within the system. The aims of the two systems are sufficiently unique, however, that we have treated them separately. In TL systems, the primary goal is to provide teachers with a vetted level for a text that corresponds to students' reading levels. The major aim of the R+E systems, which are prominently represented in today's world of the CCSS (Appendix A, 2010), is to involve teachers in identifying text features that can promote (or impede) students' capacities to read a text, rather than assigning a specific level to a text. Text mapping, unlike the other two qualitative approaches, focuses less on syntactic or lexical complexity and more on conceptual complexity, with its emphasis on clearly describing the logical relationships among ideas in a text. In terms of purpose, it is closer to the R+E systems than the TL systems because, in specifying the relations among key ideas in the text, it has direct implications for supporting comprehension through instructional scaffolds.

The TL systems suggest who ought to be able to read a particular book, either on their own or with help; R+E and TM systems indicate the scaffolds and supports a teacher might need to provide in a given classroom to help a range of students work their way through the text. Another way to characterize the distinction is that the TL systems are more text-centric, while the R+E systems are more reader-centric in their end goals. The TM approaches lie somewhere in between; they emphasize key ideas in the text but from a conceptual rather than a structural perspective. These distinct purposes are important to keep in mind as we review these three general approaches to the qualitative analysis of text complexity.

Text Leveling

The leveling of texts by expert judges is not a recent phenomenon (see, e.g., Carver, 1976; Singer, 1975). However, this procedure was not prominent until readability formulas were downplayed as a criterion for textbook selection in America's largest states (California English/Language Arts Committee, 1987; Texas Education Agency, 1990). The Reading Recovery levels (Peterson, 1991) that have evolved into guided reading levels (Fountas & Pinnell, 1996, 2009) were a response to this need.

Reading Recovery and Guided Reading Levels

The first systematic attempt at implementing a wide-scale text-leveling scheme emanated from Peterson's (1991) dissertation research at Ohio State University on

RR books. Peterson started with the books that were in use as exemplars of reading recovery (RR) levels in New Zealand at that time. The degree to which Peterson's work resulted in changes in assigned RR levels is not certain, but she did identify four criteria that distinguished among books judged to be at 20 different levels spanning the reading acquisition period: (1) book and print features; (2) content, themes, and ideas; (3) text structure; and (4) language and literary elements. Descriptions were written to show how the features differed from level to level, but the features themselves were not analyzed as separate components. Sample texts that exemplify particular levels were provided, but the details of how and why these texts illustrate particular features at particular levels were not specified. In this approach, then, a set of criteria is offered, but judges assign a text holistically with the influence of particular criteria on the whole score uncertain.

Following Peterson's (1991) work on RR levels, Fountas and Pinnell (1996) applied the leveling system to texts in classrooms within the context of their approach to guided reading. Their system was similar to RR levels, although Fountas and Pinnell used a 26-level (as compared to 20-level) scale that extended to sixth grade and, subsequently, to eighth grade (Fountas & Pinnell, 2009). Their criteria for scaling books include the same content foci with the most recent iteration of the system called the F&P Text Gradient (Fountas & Pinnell, 2012) identifying 10 factors (genre, text structure, content, themes and ideas, language and literary features, sentence complexity, vocabulary, words, illustrations, and book and print features).

The Fountas and Pinnell (1996, 2009) process of evaluation is the same as in RR. A rater uses the descriptions to assign a book to a level, under the untested assumption that the steps between levels for any of the key traits changed by roughly the same amount from level to level. In essence it operates like a holistic writing rubric; that is, a judge might examine a text on several different dimensions, but then amalgamate all of that feature-by-feature information to reach a judgment that the text should be assigned to a particular level. Scores or levels are not reported for individual categories (e.g., content, text structure); instead, the different categories or scales inform the holistic rating.

No research studies have reported on the relative weight given to different dimensions in these holistic ratings or whether the dominant factors vary for different types of texts or different levels of readers. For example, print features might be expected to weigh more heavily at the early levels (i.e., A–E), but a variable like referential cohesion or syntactic simplicity might dominate at high levels (i.e., V–Z). The role of individual variables, it would seem, has been subsumed into a holistic rating. Holistic scoring may obscure between-criterion variability; it would not be hard to imagine a text that was at level T on vocabulary but only at level M on structural elements. Graesser, McNamara, and Kulikowich (2011), using more quantitative analyses for five separate linguistic elements, found that texts judged to reside at a particular level of readability can vary widely on an array of specific elements of text complexity.

Publishers and educators have applied the text leveling of RR and guided reading to literally thousands of texts. Despite its widespread use, we were unable to find any reports of reliability across coders in leveling texts for either scheme. Further, while proponents of this form of leveling present it as an alternative to readability formulas, one of the only studies of its validity (Hatcher, 2000) has reported a strong correlation ($r = .82$) between text levels within RR and the principal factors that make up traditional readability formulas (word frequency and sentence length).

AQ: 3

We could find no studies that examined how instruction with texts ordered according to either RR or guided reading levels influenced reading acquisition. We located a single study (Hoffman, Roser, Patterson, Salas, & Pennington, 2001, reviewed next) that examined student performance on texts at different RR and guided reading levels.

Scale for Text Accessibility and Support (STAS-1)

Similar to guided reading levels, the STAS-1 (Hoffman et al., 2001) uses expert judges to rate texts on dimensions of text complexity. Unlike the holistic scores of guided reading levels, levels on the STAS-1 are a product of independent ratings of several raters on two scales—decodability and predictability. Hoffman and his associates used a methodology (Carver, 1976; Singer, 1975) in which experts use anchor passages that had been previously ordered according to specific criteria. For example, on the decodability criterion, texts rated as highly decodable (1 on the scale) contain words with consonant-vowel-consonant (CVC) patterns, single syllables, and short high-frequency words, while minimally decodable texts (rated as 5) contain irregularly spelled words and a variety of vowel patterns. In between these end points are three interim points: (2) very decodable, (3) decodable, and (4) somewhat decodable. A comparable five-point scale used four predictable features (picture support, repetition, rhyming elements, and familiar events/concepts) to guide raters in making an overall rating of predictability. Hoffman et al. (2001) reported that, on the basis of 21 texts (three texts from the seven earliest of RR levels), the average correlation among judges' ratings was .78.

Hoffman et al. (2001) examined how well the STAS-1, RR levels, and guided reading levels of texts were able to predict student performance using empirical scores on measures of accuracy, fluency, and rate across three instructional conditions (text preview, word preview, and no preview). All three TL (RR, F-P, and STAS) systems yielded small to moderate—but statistically significant—correlations with accuracy, fluency, and rate metrics in the .2 to .4 range, with the consistent predictive advantage going to STAS-1 over the other two leveling systems. Significant but unsurprising effects were found for reader ability: more able readers were more accurate, fluent, and faster. Further, those students who received adult modeling in the form of a text preview or a sight-word preview achieved significantly higher levels of performance on fluency and accuracy indices than students in the "no preview" condition. Consistent with the perspective outlined in the model of Valencia et al. (2014, in this issue), scaffolding allows students to read texts that would otherwise be beyond their grasp.

The work of Hoffman et al. (2001) illustrates that particular dimensions of texts can be defined and that raters, when given clear criteria, can sort a group of texts reliably on a recognized trait of beginning reading texts, such as decodability or predictability. This is a particularly important finding for the texts of early reading on two counts. First, we know early texts (grades K–2) defy the reach of most of the quantitative systems of analyses (see Hiebert & Pearson, 2010) currently available. Second, the Hoffman et al. work is exemplary in terms of its research methods, particularly their emphasis on evaluating the reliability of scores assigned by judges using rubrics and anchor texts to assign books to levels and their attention to the concurrent and predictive validity of their scales; in fact, no other efforts to build

text-leveling systems grounded their efforts in real student performance on commonly used measures of reading.

Rubrics and Exemplars (R+E)

Over a period of nearly half a century, professionals have used two fundamental tools—rubrics and anchors—to score student writing (DiPardo, Storms, & Selland, 2011). In the rubric/anchor system, human judges identify a set of traits that characterize effective products, usually by examining artifacts that vary widely in holistic/impressionistic judgments of quality. These traits are placed on a continuum where less mature forms of the trait anchor one end and more sophisticated forms, the other. Each trait and its manifestations across the continuum are described as a rubric. The anchor metaphor is significant: Examples of student work that typify key levels or points along the continuum are often referred to as "anchor papers." The logic of the system (rubrics, along with anchoring exemplars) has been applied to a host of phenomena in which human judgment is involved in scoring or ranking performance other than in writing: debates, speeches, athletic events, and even applications to universities.

In adapting the logic of the writing rubric model to text analysis, the operative term has been exemplars rather than anchors. But procedures have been similar: identify important traits, develop descriptions that position levels of those traits along a set of continua, and locate anchor texts that typify points along those continua, or, in the case of holistic rubrics, a continuum.

The CCSS writers cited three references for their recommendations of a rubric for assessing text complexity qualitatively (ACT, 2006; Chall, Bissex, Conard, & Harris-Sharples, 1999; Hess & Biggam, 2004). Of these three, the Hess and Biggam system was designed and used for professional development in teachers' selection of texts for their students. Their scheme consisted of seven features (word difficulty and language structure, text structure and discourse style, features of genre/text type, background knowledge and/or degree of familiarity, level of reasoning, format and layout, and length of text). In revising the system based on feedback from professional development efforts, Hess and Hervey (2010) provided separate rubrics for narrative and informational texts and reduced the system from seven to five traits, with each trait presented along a continuum of simple, somewhat complex, complex, and very complex. Neither validity nor reliability information on this system has been reported, nor has this system been published. Both the QATD and ACT systems are more readily available than the Hess and Biggam work.

The Qualitative Assessment of Text Difficulty (QATD). The first qualitative system available through an academic publisher was the QATD (Chall et al., 1999). On the apparent belief that discipline matters in analyses of difficulty, Chall and her colleagues built four scales—each specific to one of the four major "content areas" in the system—literature, popular literature, science, and social studies. Each scale, all of which are summarized on the right half of Table 1, describes not the features of the texts at different levels of complexity but rather the knowledge and/or processes readers need to engage to be successful at successive levels of text difficulty. Each scale has four or five primary traits, with some overlap across scales. Each trait is unpacked for particular grade levels, as illustrated in Figure 1 for vocabulary. Each scale is also anchored by a set of benchmark texts, one for each developmental level on the scale,

T1,AQ:4

F1

Table 1. Comparison of Rubrics: Common Core State Standards (CCSS), ACT, and Qualitative Assessment of Text Difficulty (QATD)

CCSS		ACT	QATD			
Narrative Levels of meaning	Informational	Relationships	Popular fiction	Literature Skill in literary analysis	Social studies • Levels of reasoning • Cognitive density	Science • Levels of reasoning • Density of ideas
	Purpose	Purpose				
Structure	Structure	Structure	Familiarity with sentence structure	Familiarity with sentence structure	Familiarity with sentence structure	Familiarity with sentence structure
Language conventionality & clarity	Language conventionality & clarity	Vocabulary & style	Knowledge of vocabulary	Knowledge of vocabulary	Knowledge of vocabulary	Knowledge of vocabulary
Knowledge demands: life experiences; cultural/ literary knowledge	Content/discipline knowledge	Richness	• Cultural & literary knowledge • Depth & breadth of experiences	• Cultural & literary knowledge • Depth & breadth of experiences	Subject-related & general knowledge	Subject-related & cultural knowledge

QATD's Knowledge of Vocabulary for Literature

Reading Levels	1-2	3-6	7-12	13-15 (College)
	Mainly familiar words, often repeated	More varied, but generally familiar; some figurative language	Increasing number of uncommon words; nonliteral meanings	Wide vocabulary and range of meaning levels

ACT's Vocabulary

Uncomplicated	More Challenging	Complex
Familiar	Some difficult, context-dependent words	Demanding, highly context dependent

CCSS's Language Conventionality & Clarity

Literal	Figurative or ironic
Clear	Ambiguous or purposefully misleading
Contemporary, familiar	Archaic or otherwise

CURRENT STATE OF QUALITATIVE TEXT ANALYSES

	unfamiliar
Conversational	General academic and domain-specific

Figure 1. Illustration of traits in the three rubric/exemplar systems: Language Conventionality & Clarity/Vocabulary

to clarify how the scale differs across levels. The science and the social studies scales are each represented by two sets of anchor texts, reflecting concessions to both subdiscipline (life science and physical science for the sciences) and genre (narrative and expository accounts for history/social studies). This complex array of knowledge sources, processes, and exemplars reveals the concern that Chall et al. held for conceptual content, not simply linguistic features.

The major contribution of the Chall et al. approach to qualitative analysis is to remind us that not all texts demand the same sort of cognitive and linguistic processing—that subject matter demands (science vs. literature vs. history) necessarily

shape the ways in which readers engage the text, which connects the QATD system intimately with the disciplinary grounding of the CCSS.

ACT: Reading between the lines. In analyzing college readiness of high school students, ACT (2006) concluded that it was not the level of questions they were asked but the complexity of the text they were required to read that sorted students into levels of preparedness for college; in short, text mattered more than task, at least insofar as they adequately measured task performances. They identified three kinds of texts—uncomplicated, more challenging, and complex. These three levels of texts were differentiated on the basis of five traits that ACT writers described with the mnemonic RSVP: R (relationships, richness), S (structure, style), V (vocabulary), and P (purpose). In contrast to Chall et al. (1999), ACT scholars did not develop separate scales for disciplines, genres, or even broad categories like expository versus narrative texts.

The process of rubric development is not described within the ACT report, nor is any information given on the scoring/sorting—who did the scoring or what the particular ratings were on the traits that make up the rubric. Nor is evidence provided about the weighting that was assigned to particular elements of the rubric in determining the complexity of particular passages. Even so, it remains, in our estimation, the most interesting of all the qualitative systems, largely because of its commitment to close analysis of the particular features that render particular texts more or less accessible.

The ACT report does not provide exemplars per se but instead offers annotated versions of texts that represent two of the three levels (complex and more challenging texts). Each annotation begins with a summary of the critical content of the passage and goes on to describe the features of the text that account for the challenges students may experience when reading and answering questions about it. A portion of an annotation illustrating a complex text within the prose fiction category can be found in Appendix A. The goal of the annotation is to convey information about the ways in which text features influence readers' meaning-making, rather than descriptions of the text features per se. The annotation illustrates information useful for instructional decision making: which features of the text may create obstacles for students and which could be the focus of instruction that grows student capacity with a particular type of complex text. Of all the systems, this one shows the most potential to provide direction for teachers on how to scaffold texts that challenge students.

Common Core State Standards (CCSS) and its extensions. The qualitative system within the CCSS (NGA/CCSSO, App. A, 2010) is a hybrid of the qualitative systems described thus far, but it relies, by its own admission, more on the ACT system than the other two systems (Chall et al., 1999; Hess & Biggam, 2004). The rationale behind text classifications in particular grade bands, it would be expected, can be explained with elaborate annotations (as was done with the ACT system) of how the features of the rubrics either do or don't apply to particular texts. The CCSS developers did not take this route. Rather, they provided one-page evaluations (referred to as annotations), applying the tripartite complexity approach (i.e., quantitative, qualitative, and reader-task) for three of the 168 exemplar texts identified in Appendix B of the Standards—two from the grade 9–10 band and one for the grade 6–8 band. For example, the qualitative summary of *The Grapes of Wrath* consists of four succinct paragraphs describing the overall theme of this approximately 200,000-word book. No distinction is made in Steinbeck's two literary styles, one

present in chapters describing the plight of many farm families during the Depression and the other in the narrative of the Joads. The space devoted to the quantitative measure of *The Grapes of Wrath* is almost equivalent to that for the qualitative dimensions, attempting to explain why the text is not a grade 2–3 text even though its readability places it at that level. A cryptic statement is made for "local" determinations of reader-task considerations with a final section of the annotation devoted to the recommended placement. The entire point of the exercise appears to be to assign the book into a grade band and to justify this placement, rather than to provide information that might aid teachers in designing lessons that aim to support students in reading increasingly more complex texts.

Kansas system. The text-complexity model proposed within the CCSS has been adapted, extended, and applied in the Kansas Qualitative Measures Resources (Copeland, Lakin, & Shaw, 2012) in a four-step process that draws on the three recommended approaches, beginning with the quantitative, moving to the qualitative (with four rather than two levels of the traits), advising reviewers to attend to reader and task considerations, and culminating in a recommendation for placement in the appropriate text-complexity band of the CCSS. This model appears to have generated considerable interest. For example, the Model Content Frameworks developed by the Partnership for Assessment of Readiness for College and Careers (PARCC; PARCC, 2012) assessment consortium has suggested a similar procedure, as have several states (see, e.g., Georgia Department of Education, 2012).

Achieve the core qualitative rubric. A second adaptation of the qualitative rubric of the CCSS has been added to the Achievethecore.com website, the resource site for Student Achievement Partners (2012), the agency that held the contract for the writing of the CCSS. The rubric itself is similar to the one in Appendix A of the Standards (NGA/CCSSO, 2010, p. 6). The presentation, however, has been modified to include a place for reviewers to identify which trait trumped the others in a judge's decision to place the text in a given grade band. Reviewers are also asked to assign an instructional and an independent level to the text. The emphasis is on the placement to ensure designation of a single level, not on the content to be taught or the unique challenges of a given text.

Text Maps

Text maps depart radically from both text-leveling (TL) and rubric plus exemplars (R+E) systems. In text maps (TM), the focus is on the conceptual structure of the text; for either narratives or informational texts, the result of text mapping is a diagram of the text. For stories, it most often resembles a flow chart of the sort popular in story mapping (e.g., Pearson, 1984) and story grammar analyses (e.g., Stein & Glenn, 1979). For informational texts, the diagrams tend to be elaborate and complex, with multiple nodes and branches representing the networks of ideas within content-area texts (see, e.g., Armbruster, Anderson, & Ostertag, 1987).

Text mapping has been used within the NAEP since the late 1980s to ensure that texts have sufficient conceptual grist for inclusion in the assessment and that the items developed for NAEP passages assess important content and focus on the higher-level nodes in these elaborate semantic networks. As we have found for other qualitative systems described in this review, this procedure has not been examined in

enough detail and with enough scrutiny to have yielded analyses that have found their way into peer-reviewed research outlets.

The specifications for the procedure, however, are extensive (American Institutes for Research, 2008). Internal documentation of the procedure by contractors is presumably extensive as well, although such documentation could not be obtained for this article. The NAEP appears to have used text maps in item creation since the 1992 NAEP (National Assessment Governing Board, 1991), after the successful experiences of two states, Illinois and Michigan, in using these maps for their state assessments had been reported (Valencia et al., 1989; Wixson et al., 1987).

The essential move in text mapping is examining the ideational structure of the text by focusing on the key ideas and displaying them visually in a graphic that highlights the relationships among those key ideas. Protocols for literary and informational texts are different because of differences in these text types. Evaluators discuss their maps with one another at key points to ensure fidelity in representing key ideas and relationships. Discussion occurs before item development as well as after to ensure fidelity between the maps and the items as well as with scoring (rubrics) procedures for short- and extended-constructed responses.

AQ: 5

For both narrative and nonnarrative texts, mapping begins with a thorough reading of the text, followed by summarizing the selection's theme (narrative) or purpose (nonnarrative). After the shared processes of reading the text and writing a concise but comprehensive summary of theme/purpose, the protocols for narrative and nonnarrative take different forms, reflecting the different content of the two text types.

Narrative maps are used for literary texts with plots (i.e., some form of problem, conflict, resolution), including tales, mysteries, and realistic and historical fiction. The protocol for the narrative map captures the structure of fiction—themes, plot structure, setting, characters, and author's craft (portions of a typical narrative map appear in App. B). The process begins with identifying themes at both the story level (specific events of the narrative) and abstract level (general concepts that run through the narrative). The interrelatedness of text features is emphasized, such as the manner in which setting or the roles of characters influence plot.

Nonnarrative maps are used for texts such as speeches, exposition, descriptions, explanations, argumentative essays, and other documents. Nonnarrative maps are supposed to capture the hierarchical organization of information, with multiple levels of ideas (central, major, and supporting). Where possible and appropriate, the maps also identify the role of text features (e.g., subheadings, charts, and illustrations) and elements of the author's craft (e.g., figurative language and rhetorical devices).

A nonnarrative text map (an example of which appears in App. C) begins with the central idea and purpose and then maps out major and supporting ideas and role in text organization. An organizational element, such as comparison structure, might be highlighted, after which the major and supporting components (what is being compared and on what criteria) for the element are depicted hierarchically in a portion of the map. Like other qualitative approaches to analyzing text structure, mapping systems employ criteria, rubrics, and exemplars to train researchers to create maps and use interjudge agreements to examine the reliability and validity of the text maps they construct.

Summary

We have reviewed three different types of qualitative systems and elaborated on the ways in which they serve one of two general purposes for their use: Text-leveling systems are designed exclusively to enable a better match between students' abilities and the texts we ask them to read. Rubric + Exemplar systems and Text Maps highlight parts of texts that deserve special attention and/or instruction when we ask students to read and understand them. Additionally, several of the R+E systems also result in assigning a text to a level, namely, the CCSS system and its derivatives, such as the Kansas system. Most important to remember about these systems is that the research base documenting their efficacy for either of these purposes is very meager. Even so, TM systems can be very useful in identifying features or segments of text that deserve special instructional treatment.

Lingering Issues

As important as it is to employ qualitative analyses as a ballast for or complement to quantitative indicators of text complexity, it is even more important to refine our qualitative indicators and analyses so that they will be able to instill enough confidence in potential users to earn equal status alongside quantitative indicators in making decisions of consequence about which texts to use with whom and how. If any qualitative indicators are to achieve this status, we will, as a field, have to settle a number of lingering issues regarding their construct validity and implementation, among them issues of (*a*) purpose, (*b*) teacher professional development, (*c*) exemplars, and (*d*) developmental progression.

Staying True to the Purposes of Qualitative Text Analysis

In the final analysis, the question of interest about qualitative systems is, What are they good for? How can they help us in ways that quantitative systems cannot? In this article we have highlighted the two major purposes—matching students to texts and unearthing the "tricky parts" of particular texts for support during reading. In theory, if we do a good job of matching texts to students, they should be able to manage most texts without too much intervention from teachers. But if our goal is truly to up the ante in text complexity (a central tenet of the CCSS), then the second purpose of highlighting challenging features for instruction will be even more important than the matching function.

A third purpose of qualitative analysis, not discussed thus far, may be equally as important as the two avowed purposes. Qualitative analyses, both the R+E and TM systems, can serve to vet, validate, and/or adjust the recommendations of quantitative systems. Qualitative analyses will serve a critical function in ensuring that texts are assigned to appropriate levels. Qualitative analyses, for example, will prevent us from concluding that we can use *The Grapes of Wrath* in grades 2–3 in spite of its measured Lexile level of 680. *The Grapes of Wrath* has content that will challenge many of the grade 9–10 students who are expected to read it. The function of the qualitative scheme of the Common Core and the various spin-offs (e.g., Kansas, Achieve the Core) appears to be to provide a second sorting score. In all of these endeavors, qualitative analyses are used to vet the appropriateness of a quantitative assignment.

The measurement issues with quantitative systems that rely on syntax and vocabulary have long been documented (Anderson, Hiebert, Scott, & Wilkinson, 1985; Klare, 1984). As a general rule, quantitative systems tend to underestimate the complexity of narrative texts (e.g., short sentences typical of dialogue lower readability scores) and overestimate the difficulty of informational texts (e.g., repetition of rare, technical vocabulary raises readability scores) (Hiebert, 2011).

Supporting instruction for challenging texts. When the purpose of qualitative systems is to support instruction, the focus on ensuring that texts are sorted into appropriate "fifth-grade" or "eighth-grade" bins is less compelling than providing guidance for teachers in implementing lessons that provide students with scaffolds and skills for navigating texts that are just out of their reach. The ACT annotation and the NAEP text maps provide precisely this sort of guidance (see Apps. B and C). By contrast, the application of the R+E of the CCSS (App. A) provides little guidance for instruction. A missed opportunity is Steinbeck's (1939/2006) use of mixed genres in *TGrapes of Wrath*, in which he weaves the content from previously written articles on the conditions of farmworkers into the rich narrative of the Joad family. It isn't that such opportunities could not be included in the CCSS guidance; it is rather that the examples we have been provided thus far don't offer that level of specificity.

Support for teachers in teaching and selecting texts

If qualitative indicators of complexity are to support improvements in students' comprehension of challenging text, they will first have to influence teacher beliefs, knowledge, and, ultimately, practices. Teachers who don't know why some characteristics of text, some purposes for reading, some comprehension tasks are harder than others will not be in a position to select texts that are likely to "hit the just right mark" for particular individuals or groups. And without this knowledge, they certainly won't be able to offer scaffolding that allows students to access the key ideas from text that are just beyond students' reach. This means that professional learning, and hence professional development, is a key to increasing the salience and influence of qualitative schemes for analyzing text complexity.

Surely the level of information required of teachers will differ as a function of age of the readers and a text's developmental complexity. A teacher working with a class of eighth or ninth graders on *The Book Thief* (Zusak, 2007) will presumably need different information about text, task, or knowledge complexity than a second-grade teacher working with students on *The Treasure* (Shulevitz, 1986). To appropriately teach the latter, an understanding of parables (*The Treasure*) is useful, as is an understanding of critical concepts (e.g., *inscription* in *The Treasure*). However, the level of information required to work with students on *The Book Thief*—especially students whose knowledge of the Holocaust is limited—will be extensive.

Especially critical is the question of whether teachers need to do these rich qualitative analyses themselves or whether there are ways in which teacher collectives and/or publishers can provide some of the information. Even if publishers provide the information, teachers will need to engage in in-depth analyses of complex texts at particular levels in published anthologies in order to satisfy themselves that their authors of the teacher editions "got it right." As a practical concern, a question we will have to answer is whether teachers in the role of "reading coach" for schools can give the kinds of supports required.

Even at the beginning levels, it is doubtful that an overall designation of "uncomplicated" (ACT), an alphabetic letter on a scale of A through Z (Fountas & Pinnell, 2012), complex (CCSS), or grade 3 level reader (Chall et al., 1999) will aid teachers in providing the instruction required for a truly complex text. Presumably, a text that is truly complex for readers requires the kind of scaffolded coaching that has been described as part of deliberate practice (Ericsson, 1996). That is, there is something for learners to learn, and the teacher must do what is required to help them dig it out. Generic ratings (e.g., Level A, moderately complicated, requires grade 3 skills) will be inadequate to provide the kind of instruction that grows students' capacity to read progressively more complex texts across the grades—the essence and explicit goal of Standard 10 of the CCSS. Of course, it is not really the purpose of such general ratings to provide that level of specificity. But then, whose job is it to provide such guidance? True, interpreting the qualitative (and quantitative) information in relation to readers and the task is the teacher's milieu. But teachers can benefit from suggestions and frameworks. Every teacher should not be responsible for discovering salient features of particular texts that are likely to challenge readers at particular developmental levels. One form of potential guidance can be in the form of exemplar texts, which teachers can use in studying specific texts.

The Tyranny of the Exemplar

Previously, we have pointed to the role of exemplars in the ACT system (ACT, 2006), Chall et al.'s (1999) QATD, and various text-leveling systems (Fountas & Pinnell, 1996, 2009; Hoffman et al., 2001; Peterson, 1991). Exemplars are the concrete realization of the phenomenon being judged, and they make abstract rubrics come alive so that judges know what that phenomenon looks like when they see it. A common characteristic of these various ways of addressing complexity is that the creators of each system develop and implement some sort of vetting standards for determining where texts "fit" in their particular complexity continuum. The vetting is typically carried out by trained professionals, who use their deep experience with texts and readers along with specific criteria for selecting exemplars that they acquire in some sort of training procedure.

But the exemplars in the CCSS, both in the Standards themselves and also the Standards' Appendix B, present dilemmas that do not surface in the various text level/complexity systems. The basic difference is that exemplars in a policy document play a different role than in a technical document that describes a procedure for establishing text complexity.

Canonical texts. First, protestations to the contrary (e.g., these examples are meant to illustrate the range of types of text that might be used in a school reading program), exemplars often get interpreted as a canon. So instead of illustrating the sorts and range of texts that might be used, the exemplars become the entire population that educators use in a grade-level band. In short, the exemplars become the canon of texts that are taught. We have labeled this dilemma the "tyranny of the exemplar" because it is hard for any of us to resist believing that if a text is good enough to exemplify a level, then it ought to be taught at that level. And, indeed, some of the materials currently under development suggest that the exemplars provided in the CCSS are making their way into curriculum packages (EngageNY, 2013). But this is a temptation that must be resisted lest we marginalize all attempts by educators to

adapt the portfolio of texts used in specific district and school settings to the needs and interests of their students.

This aspect of the ELA standards—the subtle transformation from exemplars into canon—is most strident in terms of state autonomy. The standards promise in the introduction that states, districts, even teachers will have autonomy in curricular choices, but the dominance of the exemplars betrays such a promise; the list will become the canon unless some dramatic pronouncement is made or some significant step is taken. Since the standards say that the exemplars are only illustrative of the range, the step must be bold. Of this we can be sure: the smaller the list, the more likely it will become a canon. So one useful step might be to expand the list so dramatically that no district or school could possibly cover all the exemplars. Another might be to require states and districts to develop their own lists, perhaps even contributing them to a national exemplar bank. A third might be to establish a commission that every year has the task of adding newly published works to the exemplar bank. Try as the standards might to deny their canonical role, it is the default role they will serve unless specific steps are taken to rein in that natural tendency. Only a widespread concerted effort with strong policy support will prevent an unintended canon of exemplar texts.

Unwarranted assumptions of homogeneity. Second, at least through grade 5, the use of "bands" that are considered more or less homogeneous is problematic. While it might make sense to have an internally undifferentiated band that defines the range of texts that a typical high school junior or senior can read, it does not make sense to lump texts for grades 2 and 3 into an undifferentiated band. Here's the issue: Relative to one's starting point, the proportion of intellectual growth from the beginning of second grade to the end of third grade is much greater than the comparable proportion of growth from the beginning of eleventh to the end of twelfth grade. In the earlier grades, dumping a set of texts into a grade band without specifying where in the grade band students would be expected to read any given text leads to confusion and even unreasonable expectations for our youngest and most vulnerable readers. By suggesting, perhaps even mandating, that students in the first grade of a band should be able to read the most complex of texts within that band with guidance (see, e.g., Standard 10 for literary text, grade 2, NGA/CCSSO, 2010, p. 13), we end up with unrealistic expectations for at least some of the students in the band. By the way, it is exactly this sort of problem that well-articulated and well-validated qualitative analyses must be able to solve. Quantitative approaches don't have the capacity to evaluate these deep knowledge demands that don't emerge in a surface analysis of complexity.

The vetting problem. A final problem with exemplar texts in the CCSS is that the Standards document provides no account of how text band assignments were made. The document requires users to exercise blind faith in an undocumented process. With so much at stake, namely, the well-being and academic progress of our children, procedures that can be subjected to scientific scrutiny rather than blind faith are a more appropriate standard for fixing the expected levels of difficulty of text.

Rethinking Developmental Progressions

As with any framework designed to promote, examine, and monitor student learning, the question of what develops over time and across grade levels is critical to

the CCSS. Such theories of development are always implicit—but usually explicit—in documents that guide learning and teaching, and the CCSS document is no exception. Thus the first question for our consideration is, What is the theory of development underlying the CCSS? The second—Did we get it right, or right enough at least so that if we enact the CCSS we will promote student learning and the ability to handle the range of texts that our schools and society require of each generation of citizens?

Implicit or explicit progressions? If one looks at the reading standards themselves, there is an attempt to build an explicit theory of task development—what we ask students to do from one level to the next. Unfortunately, the progressions offered are more ad hoc than systematic, let alone theoretical, in delivery. As Pearson (2013) and Applebee (2013) have noted, the changes in focus (what the reader is asked to do in the name of the standard), scope (how much text the reader would have to consult to complete the task), and support (what sorts of scaffolds are present to help the reader carry out the task) vary considerably in what seem like random ways across the bands of grade level for which specific iterations of the standards play out. The net result is that one is baffled about why, for example, analogies and allusions first appear in Standard 4 (vocabulary usage) at grade 8 and are gone by grade 9. Are we to infer that grade 8 is the first point in the curriculum at which they can or should be addressed? Or that they should not be continued in grade 9? Similar discontinuities abound at every level of the standards (see Pearson, 2013, for more examples in grades K–5).

Other things besides tasks also develop in the CCSS, namely, both the structural and the conceptual complexity of the texts encountered. And these changes constitute an answer to the question, Why does reading become more challenging as students move from one grade level to the next?

What changes occur in text features across the developmental progression? The rubrics of the CCSS, ACT, and QATD (see Fig. 1) aim to answer this question. All three of the systems share the trait of vocabulary (illustrated in Fig. 1), structure (although the QATD focuses on the structure of sentences while the other two focus on text structure as well), and knowledge demands. There is somewhat more ambiguity in terms of levels of meaning or relationships among ideas and literary analysis (QATD), but presumably this trait represents the degree of inference required to construct meaning.

Excerpts from narrative exemplars from the beginning, middle, and end of the CCSS's staircase of complexity (Table 2) illustrate the challenge of ferreting out the implicit theory of text-complexity development across levels. With respect to text structure, a surface-level examination suggests the texts are not substantially different from one another, but the texts vary considerably on other variables, as the subsequent discussion illustrates.

T2

Knowledge demands. When it comes to knowledge demands, there are transparent differences. The overt decision making of an individual to commit a crime in *Crime and Punishment* is likely more demanding than understanding the squabbling between siblings in *Little Women* or deciding what should be planted in *The Stories Julian Tells*. Reading about deliberately choosing to commit a crime is inappropriate for primary-level students. How much "harder" it is to understand planting, sibling squabbling, or details of a plan to commit a crime is less certain.

Table 2. Excerpts from Exemplars of Narrative Texts for Grade Bands

Text	Grade Band	Excerpt
The Stories Julian Tells (Cameron, 1981)	2–3	My mother gave Huey and me baths. She said we were darker than the garden. She said we were dirty enough that she could grow plants on our hands and knees.
Little Women (Alcott, 2008)	6–8	"I'm not! And if turning up my hair makes me one, I'll wear it in two tails till I'm twenty," cried Jo, pulling off her net, and shaking down a chestnut mane.
Crime and Punishment (Dostoyevsky, 1996)	11–CCR	Yes, my hat is too noticeable. It looks absurd and that makes it noticeable. With my rags I ought to wear a cap, any sort of old pancake, but not this grotesque thing.

Sheer length. One feature of texts that no analysis has yet captured is their sheer length. The excerpt for grades 2–3 comes from a 1,200-word chapter of a book in which each of six chapters (i.e., 7,200-word book) tells another story from Julian's life, each based on experiences of middle-class children. The *Little Women* excerpt is from an 88,000-word text where the persistent squabbling between Jo and Amy is a secondary theme that runs throughout the book. Similarly, the excerpt for grade 11-CCR, *Crime and Punishment*, is from a book with over 203,500 words. The character's contemplation of how trifles might thwart his success as a burglar is only a small part of the retrospective contemplation in which the character engages. But the length issue, along with its implications for the attribute that some have labeled "stamina" (Greenleaf, Schoenbach, Cziko, & Mueller, 2001; Hiebert, Wilson, & Trainin, 2010; Valencia et al., 2010), remains largely uninvestigated. One thing we do know is that with longer texts, both fluency (Valencia et al., 2010) and comprehension (Hiebert et al., 2010) decrease as students move through a longer text.

Differential importance of text features across grade-level bands. A related (to the step size) issue is the question of whether different aspects of complexity do, could, or should play differentially important roles at different levels. For example, do issues of word decodability and predictability (remember the work of Hoffman et al., 2001) matter more than syntax at K–1, while syntax matters more in intermediate grades, and yet another factor, such as levels of meaning or purpose, in middle-school texts? We suspect they do. As we learn more about the empirical development of students' capacity to cope with increasingly challenging texts, we will certainly develop insights about which facets of complexity matter most in different grade bands.

Disentangling natural covariation among aspects of complexity. In the introduction to this special issue, we raised the question of whether readability causes or merely reflects comprehension difficulty, pointing to research suggesting that sometimes more complex words and syntax may simply reflect the communication of more complex ideas (Davison & Kantor, 1982; McNamara, Kintsch, Songer, & Kintsch, 1996; Pearson, 1974–1975). For complicated ideas, there may be a lower limit on how simply they may be expressed. When it comes to reviewing complex texts for potential instructional scaffolding, teachers might be well advised to focus on the complexity of the content rather than the obscurity of the words or the syntax. Figuring out what explanations, analogies, and examples might help students negotiate tough content may be more productive than addressing rare syntax or rare words. One possible approach would be to analyze how and why an author's choices of words and syntax were just right for communicating the ideas conveyed in the text.

Of course, we do not have evidence to support this approach, but it is certainly worth exploring experimentally. And it might have the side benefit of preventing us from some very unproductive ventures into teaching syntactic complexity or drilling students on the meanings of rare words.

Mapping task complexity onto text complexity. One final perspective on developmental progressions pertains to the role of task complexity (see Valencia et al., 2014, in this issue). Task complexity is the one variable that is *not* present in any of the qualitative analyses of complexity. No one seems to have addressed the question of what students do to demonstrate their understanding of a text. In readability studies, researchers seldom specify the outcome measure that serves as the criterion, implying that any one task is just as good as any other for validating readability formulas. However, a prima facie analysis suggests that task has to matter: Asking middle-school students to identify the topic of a chapter out of a high school life science text is likely easier than asking them to critique E. B. White's (1952) use of symbolism in *Charlotte's Web*. Moreover, task must also vary at least partially independent of text; that is, one can construct a relatively simple task about a very difficult text or a relatively difficult task about a simpler text.

What is it then that makes most tasks difficult for complex texts and most easy for simple texts? Our claim is that it is the ideas themselves that drive complexity. For the most part, both the structural apparatus within which they are communicated (the sentence syntax and rhetorical frames) and the tasks we ask students to complete in demonstrating their comprehension (finding the main idea, inferring character motives, connecting ideas across paragraphs, creating a summary or a synopsis) are driven by those ideas.

The match between content and structure or content and tasks isn't perfect, and it's the imperfections that tell us that *The Grapes of Wrath* is inappropriate for the grade 2–3 band, as is *Charlotte's Web* for early grade 2 rather than in its usual grade 4 placement. But in general, harder content will come packaged with bigger, less common words; longer, more complex sentences; and more intricate rhetorical frames. Moreover, finding the main idea or inferring character motives will, in general but not always, be harder for *Crime and Punishment* than for *Stories Julian Tells*.

Notice, also, that if we are right about the centrality of content, then all of the tortured machinations about which version of a particular standard should prevail for narratives in third versus fourth versus fifth grade are unnecessary. We might be just as well off (perhaps better off) to accept the appropriateness and necessity of each of the nine anchor standards as representing the full range of tasks we'd like all students to engage in as they make their way through texts at each and every level from K through 12; then we could figure out how to find ways to embed them in the texts we decide to use at different grade levels. In short, we should let the content—the ideas—drive our placement of texts and the tasks we generate to ensure and assess comprehension of those texts. Surely, we will attend also to the structures in which those ideas travel and to the tasks we use to engage students in conversations about the texts, but we will always start and end with the ideas as the object of our analyses.

A focus on content will impose two additional requirements. First, qualitative analysis will necessarily trump quantitative analyses of texts, and, second, the analyses we engage in for structure and task will be focused on the goal of making texts more accessible to the broadest possible range of students. Such a focus will also put

quantitative analyses in proper perspective, for we will recognize that the key elements of quantitative inquiry—long words and complex syntax—are as likely to be symptoms as causes of challenging content. Armed with that knowledge, we will be better positioned to figure out how to help students manage that content, which is our most important job as teachers. This brings us full circle to one of the central goals of the CCSS for English language arts, which is eloquently stated in the introduction to the Standards, when they assert that readers who meet these standards "actively seek the wide, deep, and thoughtful engagement with high-quality literary and informational texts that builds knowledge, enlarges experience, and broadens worldviews" (NGA/CCSSO, 2010, p. 3).

AQ: 7

Appendix A

Annotation of a Complex Text Representing Prose Fiction: ACT (2006, p. 18)

This text describes two complex, well-developed characters, Sunday and Delta, and their strained yet loving relationship. One factor that contributes to the complexity of the text is its structure: the third-person narrator presents the two sisters both as they see themselves and how each sees the other.

PROSE FICTION: This passage is adapted form the novel *Night Water* by Helen Elaine Lee (© 1996 by Helen Elaine Lee).

There had been no words for naming when she was born. She was "Girl Owens" on the stamped paper that certified her birthday, and at home, she had just been "Sister," that was all. When asked to decide at six, what she would be called, she had chosen "Sunday," the time of voices, lifted in praise.

That was one piece of the story, but other parts had gone unspoken, and some had been buried, but were not at rest. She was headed back to claim them, as she had taken her name.

VOCABULARY: Beginning with the opening sentence—"There had been no words for naming when she was born"—the text uses fairly sophisticated syntax.

Appendix B

Portion of a Narrative Text Map for Eighth Grade: "Thank you, Ma'am" (from 2009 NAEP Reading Assessment & Item Specifications)

STORY LEVEL THEME: A woman's tough, but sympathetic, response to a teenage boy who tries to steal her purse causes the boy to change his behavior/attitude.

ABSTRACT THEME: Kindness, trust, and generosity are used to teach a young boy a lesson about right and wrong.

PLOT:
Problem: Roger attempts to steal Mrs. Jones' purse in hopes of getting money to buy a pair of shoes he cannot afford to purchase.

Conflict: Will Roger run or will he let Mrs. Jones help him?
Resolution: Roger reciprocates the trust and caring demonstrated by Mrs. Jones, and is given a chance to change his life.

SETTING (and how it is connected to the themes and significant ideas in the text): Urban area and small apartment where everything is in view provide a woman with an opportunity to help a young boy to see the wrongness of his actions.

CHARACTER/S* (traits that are connected to significant ideas in the text):
Mrs. Luella Bates Washington Jones/Woman

- Trusting—she leaves her purse where the boy could take it if he wanted to; provides him with a choice about going to the store with her money to buy food or eating what she has on hand
- Honest—she is straightforward with the boy and never tries to deceive him
- Caring—she does not turn him over to the police, gives him food and money

MAJOR EVENTS:**

1. Roger attempts to steal a purse of an older woman but is thwarted in his attempt by a woman who is not easily taken advantage of.
2. The woman quickly establishes her physical and emotional control over the boy.
3. She is able to judge the character of the boy and use her insights and experience to build trust between them.

AUTHOR'S CRAFT:
Tone: one of authority in the beginning changing to one of concern
Rhetorical devices
Use of italics
Significance of the title and use of Ma'am throughout
Use of slang diction
Use of "run" image throughout

*One of two characters included; 3 of 7 traits are listed
**3 of the 12 major events are given

Appendix C

Portion of a Nonnarrative Map for Eighth Grade: "Ellis Island" (from 2009 NAEP Reading Assessment & Item Specifications) (AIR, 2008)

CENTRAL IDEA: To provide a historical account of immigrants told in the words of immigrants who came to the U.S. through Ellis Island between 1892 and 1954

MAJOR IDEAS*

Org. Element—Description/Introduction:

Major Idea: Between 1892 and 1954, Ellis Island was the "doorway to America" for 17 million people.
Supporting Idea/s: Not everyone was welcome; "land of the free" was not so free to everyone.

Org. Element—Cause

Major Idea: Immigrants came from Europe to escape oppression/poverty and/or seek a better life.
Supporting Idea/s: First-hand accounts from a woman escaping Turkish oppression in Armenia, and a man from the Ukraine seeking opportunities offered by U.S.

TEXT FEATURES:
Subheadings, illustrations, use of italics to set off quotations from past immigrants
Illustration of "cattle-pen-like" method of processing

AUTHOR'S CRAFT:
Use of first-hand accounts to illustrate the points about the immigrant experience in general and on Ellis Island
Use of a doorway to America/doorway metaphor
*2 of 7 major ideas

Note

1. See www.fountasandpinnellleveledbooks.com; www.scholastic.com/bookwizard, for example.

References

ACT. (2006). *Reading between the lines: What the ACT reveals about college readiness in reading*. Iowa City: Author.

Alcott, L. M. (2008). *Little women*. New York: Puffin.

American Institutes for Research. (2008). *Reading assessment and item specifications for the 2009 National Assessment of Educational Progress*. Washington, DC: Author.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading*. Champaign, IL: Center for the Study of Reading.

Applebee, A. (2013). Common Core State Standards: The promise and the peril in a national palimpsest. *English Journal*, **103**(1), 25–33.

Armbruster, B., Anderson, T., & Ostertag, J. (1987). Does text structure/summarization instruction facilitate learning from expository text? *Reading Research Quarterly*, **22**, 331–346.

Beaver, J. (2003). *Developmental reading assessment*. Parsippany, NJ: Celebration.

Betts, E. (1946). *Foundations of reading instruction*. New York: American Book.

California English/Language Arts Committee. (1987). *English-language arts framework for California public schools (kindergarten through grade twelve)*. Sacramento: California Department of Education.

Cameron, A. (1981). *The stories Julian tells*. New York: Random House Books for Young Readers.

Carver, R. P. (1976). Measuring prose difficulty using the Rauding Scale. *Reading Research Quarterly*, **11**, 660–685.

Chall, J. S., Bissex G., Conard, S., & Harris-Sharples, S. (1999). *Qualitative assessment of text difficulty*. Brookline, MA: Brookline Publishers.

Copeland, M., Lakin, J., & Shaw, K. (January 26, 2012). *Text complexity and the Kansas Common Core Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*. Retrieved from http://www.ccsso.org/Resources/Digital_Resources/The_Common_Core_State_Standards_Supporting_Districts_and_Teachers_with_Text_Complexity.html

Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative measurement of text difficulty: What's the use? *Elementary School Journal*, **115**(2).

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, **17**(2), 187–208.

DiPardo, A., Storms, B. A., & Selland, M. (2011). Seeing voices: Assessing writerly stance in the NWP Analytic Writing Continuum. *Assessing Writing*, **16**(3), 170–188.

Dostoyevsky, F. (1996). *Crime and punishment*. New York: Bantam Classics.

EngageNY. (2013). *Lincoln Gettysburg Address*. Retrieved from http://www.docstoc.com/docs/101478560/Lincoln-Gettysburg-Address-EngageNY

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 1–50). Mahwah, NJ: Erlbaum.

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (2009). *The Fountas & Pinnell leveled book list, K–8+*, print version. Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (2012). *The F & P Text Level Gradient: Revision to recommended grade-level goals*. Portsmouth, NH: Heinemann. Retrieved from http://www.heinemann.com/fountasandpinnell/pdfs/WhitePaperTextGrad.pdf

Georgia Department of Education. (2012). *Common Core Georgia performance standards text complexity rubric*. Retrieved from https://www.georgiastandards.org/Common-Core/Documents/9%20-11%20rubrics.pdf?

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, **40**, 223–234.

Greenleaf, C., Schoenbach, R., Cziko, C., & Mueller, F. (2001). Apprenticing adolescent readers to academic literacy. *Harvard Educational Review*, **71**(1), 79–129.

Hatcher, P. J. (2000). Predictors of Reading Recovery book levels. *Journal of Research in Reading*, **23**, 67–77.

Hess, K., & Biggam, S. (2004). *A discussion of "increasing text complexity."* Published by the New Hampshire, Rhode Island, and Vermont Departments of Education as part of the New England Common Assessment Program (NECAP). Retrieved from www.nciea.org/publications/TextComplexity_KH05.pdf

Hess, K., & Hervey, S. (2010). *Local assessment toolkit: Tools for examining text complexity*. Dover, NH: National Center for the Improvement of Educational Assessment.

Hiebert, E. H. (2011). Beyond single readability measures: Using multiple sources of information in establishing text complexity. *Journal of Education*, **191**(2), 33–42.

Hiebert, E. H., & Pearson, P. D. (2010). *An examination of current text difficulty indices with early reading texts* (Reading Research Report No. 10.01). Santa Cruz, CA: TextProject. Retrieved from http://textproject.org/assets/publications/TextProject_RRR-10.01_Text-Difficulty-Indices.pdf

Hiebert, E. H., Wilson, K. M. & Trainin, G. (2010). Are students really reading in independent reading contexts? An examination of comprehension-based silent reading rate. In E. H. Hiebert & D. Ray Reutzel (Eds.), *Revisiting silent reading: New directions for teachers and researchers*. Newark, DE. IRA.

Hirsch, E. D., Jr. (Ed.). (2005). *What your fourth grader needs to know: Fundamentals of a good fourth-grade education (core knowledge)*. New York: Dell.

Hoffman, J., Roser, N., Patterson, E., Salas, R., & Pennington, J. (2001). Text leveling and little books in first-grade reading. *Journal of Literacy Research*, **33**, 507–528.

Klare, G. (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 681–744). New York: Longman.

Leslie, L., & Caldwell, J. S. (2010). *Qualitative reading inventory* (5th ed.). Boston: Pearson.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, **14**(1), 1–43.

National Assessment Governing Board. (1991). *Reading framework for the 1992 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects with Appendices A–C*. Washington, DC: Authors.

Partnership for Assessment of Readiness for College and Careers. (2012). *Model content frameworks*. Retrieved from http://www.parcconline.org/parcc-model-content-frameworks

Pearson, P. D. (1974–1975). The effects of grammatical complexity on children's comprehension, recall and conception of semantic relations. *Reading Research Quarterly*, **10**, 155–192.

Pearson, P. D. (1984). Asking questions about stories. In A. J. Harris & E. R. Sipay (Eds.), *Readings in reading instruction* (3rd ed.). New York: Longman.

Pearson, P. D. (2013). Research foundations of the Common Core State Standards in English language arts. In S. Neuman & L. Gambrell (Eds.), *Quality reading instruction in the age of Common Core State Standards* (pp. 237–262). Newark, DE: International Reading Association.

Peterson, B. L. (1991). Selecting books for beginning readers: Children's literature suitable for young readers. In D. E. DeFord, C. A. Lyons, & G. S. Pinnell (Eds.), *Bridges to literacy: Learning from Reading Recovery* (pp. 119–147). Portsmouth, NH: Heinemann.

Pikulski, J. J., & Shanahan, T. (Eds.). (1982). *Approaches to the informal evaluation of reading*. Newark, DE: International Reading Association.

Rog, L. J., & Burton, W. (2001). Matching texts and readers: Leveling early reading materials for assessment and instruction. *Reading Teacher*, **55**, 348–356.

Sherman, L. A. (1893). *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn.

Shulevitz, U. (1986). *The treasure*. New York: Square Fish.

Singer, H. (1975). The SEER technique: A non-computational procedure for quickly estimating readability level. *Journal of Reading Behavior*, **7**, 255–267.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *Multidisciplinary approaches to discourse comprehension*. Hillsdale, NJ: Ablex.

Steinbeck, J. (1939/2006). *The grapes of wrath*. New York: Penguin Classics.

Student Achievement Partners. (2012). *Qualitative dimensions of text complexity chart: 2nd–3rd grade band*. New York: Author. Retrieved from www.achievethecore.org/steal-these-tools/text-complexity/qualitative-measures

Texas Education Agency. (1990). *Proclamation of the State Board of Education advertising for bids on textbooks*. Austin, TX: Author.

Trelease, J. (2006). *The read-aloud handbook* (6th ed.). New York: Penguin.

Valencia, S. W., Pearson, P. D., Peters, C. W., & Wixson, K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership*, **46**, 57–63.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, **45**(3), 270–291.

Valencia, S. W, Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal*, **115**(2).

White, E. B. (1952). *Charlotte's web*. New York: HarperCollins.

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2014). Student reading growth illuminates the Common Core text-complexity standard: Raising both bars, *Elementary School Journal*, **115**(2).

Wixson, K., Peters, C., Weber, E., & Roeber, E. (1987). New directions in statewide reading assessment. *Reading Teacher*, **40**, 749–754.

Zusak, M. (2007). *The book thief*. New York: Alfred A. Knopf.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 1

AQ1—Per ESJ practice, I have placed the literature references in with the rest of the references, rather than having a separate list.

AQ2—I rewrote the beginning of this sentence in order to avoid starting a sentence with an abbreviation, which is against ESJ style. See also the beginning of the next paragraph.

AQ3—ESJ uses Webster's for spelling. Webster's preferred plural is formulas rather than formulae.

AQ4—Please check Table 1 carefully to make sure it's typeset correctly. It was difficult to tell how cells and columns should line up.

AQ5—ESJ closes up (i.e., does not hyphenate) words that begin with prefixes such as "non," "pre," and so on. Hence "nonnarative."

AQ6—This was originally called Table 3. I assume that you mean Table 2 (there is no Table 3 included with the manuscript). OK?

AQ7—I tried not to change the text of the appendixes. But I did add periods at the end of complete sentences (for consistency) and fix spelling. Also, in Appendix B I changed "M'am" to "Ma'am" because that's how it's spelled in the Langston Hughes text. There may have been other minor changes to regularize the text for better presentation.